The Perils of Parfit 1: Credible Commitments

On the advice of several people, I started reading Derek Parfit's *Reasons and Persons*. I haven't gotten very far, but it seems to me to be a horribly muddled book, wrong on just about every point. So perhaps this will be an ongoing series where I debunk the book in sequence. I apologize to my readers if these posts seem obvious and not very interesting. That's because I think the situations Parfit discusses are actually quite simple and it's only his muddled terminology that makes them seem tricky.

Let's clarify Parfit's discussion of self-defeating theories, which really comes down to a discussion about credible commitments. For simplicity, imagine there is no interest or inflation and your only goal in life is to maximize how much money you have. Thus an act is rational iff it contributes to that goal.

Case 1: There are two buttons. SUBTRACT removes $1000 from your bank account, ADD adds $5000 to your bank account. Obviously it is rational to press ADD and irrational to press SUBTRACT.

Case 2: There is one button, BOTH, which does both at the same time. Obviously it is rational to press BOTH, since it results in a net gain of $4000.

Case 3: There is a different button, DELAY, which adds $5000 to your bank account today and then removes $1000 in exactly one week. (DELAY is a weird button — to prevent you from using it twice at the same time, it stays down for the whole week and only pops back up once the $1000 is removed.) Obviously it is rational to press DELAY since it too results in a net gain of $4000.

Case 4: There are two buttons: DELAY, which is the same as before, and EVADE, which changes your bank account number so that none of the other buttons work. You can only press each button once and they have no other consequences. It is rational to press DELAY and then EVADE, for a net gain of $5000.

When is it rational to press EVADE? Only when you don't expect to be able to press DELAY ever again. (EVADE gains you at most $1000, while DELAY gains you at least $4000.) If you could press DELAY twice, would it be rational to hit EVADE after the first press? Of course not, it'd cost you at least $4000. But Parfit seems to suggest one is acting rationally irrationally by not pressing it. The notion seems nonsensical.

Case 5: Same two buttons, except after you press the DELAY button it engages a little impenetrable metal cover that physically prevents you from pressing EVADE. It's rational to press DELAY. Then it's rational to press EVADE, but that's kind of irrelevant, because it's also impossible.

Case 6: Same as 5, except it injects you with a serum that prevents you from pressing EVADE. Again, it's rational to press DELAY and then rational but impossible to press EVADE.

I don't see a big difference between these two cases, but Parfit seems to think the difference is vital.

Perhaps it's the fact that another person is involved that leads to the complications?

Case 7: Same as 4, except the $1000 goes into Bob's account and only Bob can press DELAY. Bob has the same notion of rationality as you and thus will only press DELAY if he believes you will not press EVADE. You could promise not to press it, but it would be irrational for you to keep that promise so Bob rightly does not believe it. However, it would be rational for you to engage the impenetrable cover or take the serum that prevents you from pressing EVADE.

There is no rational irrationality. Your goal of maximizing your money is not self-defeating. This all seems like the most obvious, unarguable stuff in the world. So I don't see why Parfit is so confused about it.

You should follow me on twitter here.

July 1, 2010